Federal Aviation Administration

DOT/FAA/AM-12/12
Office of Aerospace Medicine
Washington, DC 20591

Flight Attendant Work/Rest Patterns, Alertness, and Performance Assessment: Field Validation of Biomathematical Fatigue Modeling

Peter G. Roma,[1,2] Steven R. Hursh[1,2]
Andrew M. Mead,[3] Thomas E. Nesthus[3]

[1]Institutes for Behavior Resources
Baltimore, MD 21218

[2]Johns Hopkins University School of Medicine
Baltimore, MD 21218

[3]Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

September 2012

Final Report

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-12/12		

4. Title and Subtitle	5. Report Date
Flight Attendant Work/Rest Patterns, Alertness, and Performance Assessment: Field Validation of Biomathematical Fatigue Modeling	September 2012
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Roma PG,[1,2] Hursh SR,[1,2] Mead AM,[3] Nesthus TE[3]	

9. Performing Organization Name and Address		10. Work Unit No. (TRAIS)
[1]Institutes for Behavior Resources [2]Johns Hopkins U. School of Medicine Baltimore, MD 21218	[3]FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125	11. Contract or Grant No.

12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code U.S. Congress

15. Supplemental Notes

Work was accomplished under Cooperative Agreement 08-G-006 (SRH) and Contract DTFAAC-11-P-04677 (PGR)

16. Abstract

Fatigue-induced impairments in neurobehavioral performance capacity may compromise safety in 24-hr operational environments, and developing reliable and valid methods of identifying work/rest patterns that produce fatigue and undermine performance is important. One approach is the use of biomathematical modeling as a means of predicting, preventing, and mitigating fatigue-induced safety risks. The Sleep, Activity, Fatigue, and Task Effectiveness model (SAFTE; Hursh et al., 2004) is among the more mature fatigue models currently used in military, shift-work, and various transportation operations. The SAFTE model was constructed empirically, integrating classical physiological and circadian processes with task effectiveness predictions based on the scientific literature of standardized laboratory tests. SAFTE has been validated against accident risk in railroad operations; however, as with virtually all fatigue models, the extent to which variations in model predictions correspond to variations in actual performance capacity in the aviation environment is largely unknown. The present report offers a field validation of the SAFTE model using data from a broad sample of 178 aviation cabin crew from the 2009-2010 FAA Civil Aerospace Medical Institute (CAMI)-sponsored Flight Attendant Field Study (Roma et al., 2010). Data were collected daily throughout each individual's continuous 3 to 4-week study period. Objective sleep/wake patterns were determined via actigraphy. In addition, a personal digital assistant device was used to maintain an activity log documenting work schedules and locations, and neurobehavioral performance capacity was assessed via standardized 5-min Psychomotor Vigilance Tests (PVT) taken before and after each work day and sleep episode. Individual sleep, wake, and work patterns were entered into the Fatigue Avoidance Scheduling Tool (FAST) software for continuous records of Predicted Effectiveness (PVT Speed [1/Reaction Time] expressed as a % of individual optimum baseline). SAFTE-FAST performance predictions were then temporally aligned with the 10,659 valid PVT test sessions from the field study, and performance data from each session were expressed as Actual Effectiveness (same as Predicted Effectiveness), Reaction Time (RT), Speed (1/RT), Lapses (RTs>500 msec), and False Starts (FS; premature responses). Linear regression of mean PVT performances across 5% SAFTE prediction bins revealed significant correlations between SAFTE Predicted Effectiveness and PVT Actual Effectiveness (R^2=0.884, p<.001), RT (R^2=0.745, p<.01), and Lapses (R^2=0.486, p<.05). Identical analyses of the 7,533 valid PVT sessions completed while away on a multi-day work "trip" (i.e., excluding sessions while off-duty at home) revealed significant correlations between SAFTE Predicted Effectiveness and mean PVT Actual Effectiveness (R^2=0.889, p<.001), RT (R^2=0.819, p<.001), Speed (R^2=0.808, p<.001), and Lapses (R^2=0.484, p<.05). Finally, separate regression analyses of all valid Pre-Work (n=1,712) and Post-Work (n=1,934) PVT sessions revealed a significant Pre-Work correlation between SAFTE Predicted Effectiveness and mean PVT Actual Effectiveness (R^2=0.530, p<.05), and significant Post-Work correlations between SAFTE Predicted Effectiveness and mean PVT Actual Effectiveness (R^2=0.600, p<.05), RT (R^2=0.887, p<.001), Speed (R^2=0.539, p<.05), and Lapses (R^2=0.901, p<.001). Despite inherent technical limitations and issues of inter-individual variability, these results clearly support the validity of the SAFTE model for population-level prediction of fatigue-induced impairments in objective neurobehavioral performance capacity in extremely dynamic 24-hr field operations such as commercial aviation.

17. Key Words	18. Distribution Statement
Biomathematical Fatigue Modeling, Fatigue Model Validation, Fatigue, Performance Assessment, Work and Rest Patterns, Flight Attendants, Cabin Crew	Document is available to the public through the Internet: www.faa.gov/go/oamtechreports

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	17	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGMENTS

We thank all of our flight attendant participants for their time, contributions to the project, and dedication to their profession. We are grateful to Dr. Melissa Mallis for her contributions to the design of the Flight Attendant Field Study and participant training materials. We also thank Wendy Krikorian and Dr. Francine James (IBR, Inc.) for conducting informed-consent interviews, as well as Lena Dobbins (Civil Aerospace Medical Institute, CAMI) and Kali Holcomb (formerly at CAMI), and Carrie Roberts and Suzanne Thomas (Xyant Technology, Inc.) for their invaluable contributions to participant training.

We are grateful to Peter Wubbels, Josiah Sewell, Lianne Kitajima, Johnny Lam, and Glen Eguchi (Archinoetics) for their technical and logistical support, and to Dr. Bob Hienz, Ed Azigi, and Ginnie Gooden (IBR, Inc.) for their contributions to data processing. We are especially indebted to Marta Genovez and Zabecca Brinson (IBR, Inc.) for their herculean data management and processing efforts.

This research was in part required by a mandate from the U.S. Congress to study the effects of fatigue on flight attendants, and it was supported by Cooperative Agreement 08-G-006 (SRH) and Contract DTFAAC-11-P-04677 (PGR) from the Civil Aerospace Medical Institute, United States Federal Aviation Administration.

CONTENTS

Flight Attendant Work/Rest Patterns, Alertness, and Performance Assessment: Field Validation of Biomathematical Fatigue Modeling

BACKGROUND

Numerous factors can affect safety, performance, and quality of life in individuals working in 24-hr operational environments such as industrial shiftwork, military, health care, law enforcement, space exploration, and transportation. One issue of increasing importance to commercial aviation is fatigue (Avers, King, Nesthus, Thomas & Banks 2009; Mallis, Banks, & Dinges 2010; Nesthus, Schroeder, Connors, Rentmeister-Bryant, & DeRoshia 2007). Fatigue is generally defined as a state of tiredness due to prolonged wakefulness, extended work periods, and/or circadian misalignment, and is characterized by decreased alertness, impaired decision making, and diminished neurobehavioral performance capacity (Åkerstedt, 1995; Dinges, 1995). The very nature of 24-hr operational environments superimposed against human circadian physiology all but guarantees the systematic production of fatigue. As such, valid and reliable methods of predicting compromised performance capacity could be valuable as a means of preventing and mitigating fatigue-induced safety risks in applied settings.

One approach that has attracted attention in recent years is the development and application of biomathematical modeling as a means of predicting, preventing, and mitigating fatigue-induced risks. Among the more mature and well-regarded fatigue models is the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model (Hursh et al., 2004; Hursh & Van Dongen, 2010; Van Dongen, 2004). SAFTE is a predictive rather than descriptive model that incorporates several dynamic components such as a homeostatic sleep reservoir, circadian oscillator, and sleep inertia function (see Figure 1). Final cognitive/task effectiveness predictions integrate with these components but are based on the scientific research literature of well-controlled fatigue manipulations in well-controlled laboratory settings, ultimately relying on psychomotor speed (1/Reaction Time, expressed as a percentage of individual well-rested baseline) in the traditional 10-min Psychomotor Vigilance Test (PVT; Basner & Dinges, 2011) as the model's principal outcome metric. Originally developed with support from the U.S. Department of Defense, the SAFTE model has been adopted for use in a variety of operational contexts beyond the military, including rail, industrial shiftwork, and aviation.

Arguably the most critical aspect of any model is its predictive validity, which in the case of fatigue models and risk management is the extent to which predicted performance decrements correspond to adverse performance outcomes in the operational environment. In their 2008 report for the U.S. Federal Railroad Administration, Hursh and colleagues validated the SAFTE model against a database of 30-day work histories preceding 400 human factors and 1,000 non-human factors freight rail accidents. Although the model had no predictive power for non-human factors accidents, the relative risk of human factors-related accidents increased significantly during periods when SAFTE predicted fatigue-induced impairments in performance effectiveness (beginning at 90% of baseline with a linear increase in risk as predicted effectiveness decreased). Subsequent analyses of 350 human factors accidents later demonstrated that the relative economic risk (accident probability x [material damage + casualty costs]) was increased by 500% when

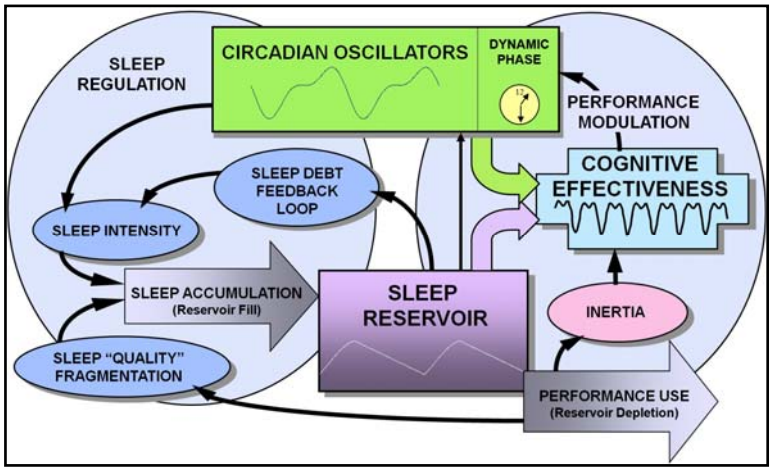

Figure 1: SAFTE Model Schematic

SAFTE-predicted effectiveness scores were at, or below 77%, whereas relative economic risk was reduced by 75% when SAFTE-predicted effectiveness was above 90% (Hursh, Fanzone & Raslear 2011). These validation data powerfully demonstrated the SAFTE model's ability to predict human factors accident risk and financial impact in rail operations; however, given those studies' retrospective design, no measures were taken quantifying changes in the engineers' neurobehavioral performance capacity underlying the accidents. Moreover, the generalizability of the SAFTE model's validity—at least the extent to which variations in predicted effectiveness correspond to variations in performance effectiveness—has never been empirically assessed within the context of commercial aviation.

To address these issues, the present report offers a validation analysis of the SAFTE model drawn from the extensive database collected during the 2009-2010 Civil Aerospace Medical Institute (CAMI)-sponsored Flight Attendant Field Study (Roma, Mallis, Hursh, Mead & Nesthus 2010). As part of a series of Congressionally-mandated projects on fatigue, a major goal of the prospective field study was to evaluate the predictive validity of the SAFTE model, using actual sleep/wake/work patterns and standardized objective neurobehavioral performance metrics taken in the "real world" by a broadly representative sample of professional cabin crew. To the best of our knowledge, the present study is the first validation of any fatigue model to use objective performance measures in the field within the extremely dynamic commercial aviation environment (cf. Civil Aviation Safety Authority [CASA], 2010; also see Spencer and Robertson, 2007).

METHOD

All human subjects procedures involved in this project were independently reviewed and approved by the Institutional Review Boards of both the U.S. Federal Aviation Administration and the Institutes for Behavior Resources. The formal letters of approval from each institution are available upon request from the authors. All data have been de-identified to protect the privacy of those who participated.

Participants

We refer the reader to Roma et al. (2010) for extensive details on recruitment, materials, and the data collection protocol for the CAMI Flight Attendant Field Study. Briefly, all eligible applicants were active U.S.-based flight attendants categorized according to three broad factors serving as the organizing framework for the study's design. These factors were Carrier Type (Network, Low-Cost, or Regional), Seniority (self-identified Senior 1/3, Mid 1/3, or Junior 1/3), and majority Flight Operations (Domestic or International). The study was designed for 210 flight attendants as shown in Figure 2. A total of 202 flight attendants completed participation in the study, and as described below, 178 individuals contributed data suitable for the modeling analysis presented herein.

Materials and Data Collection

Each participant was issued a wristwatch-shaped, water-resistant actigraphy device for continuous objective recording of sleep/wake patterns (ReadiBand™, Fatigue Science, Honolulu, HI, USA) and a touchscreen-based personal digital assistant device (PDA) for maintaining a daily activity log and collecting objective performance data

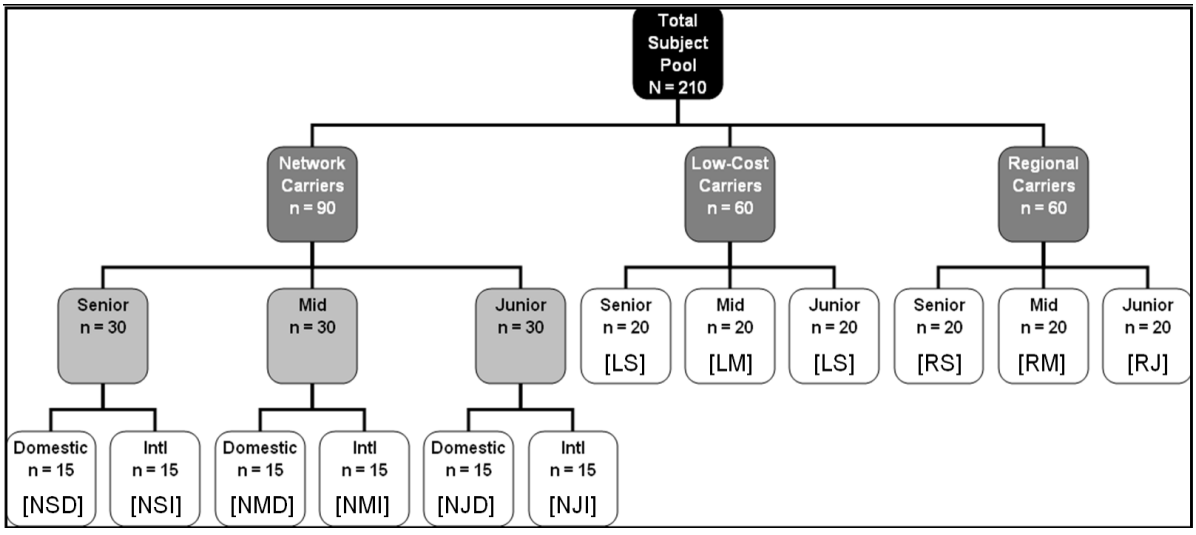

Figure 2: Stratified Field Study Design and Target Sample Sizes

(AT&T Tilt™). Using a custom-programmed graphical interface on the PDA, all participants maintained the activity log by recording the location (airport code) and local start time of various activities such as commuting, on-duty periods, off-duty periods (either at home or away on a work "trip"), and sleep episodes.

Participants were also required to complete up to four discrete test sessions per day: Pre-Sleep, Post-Sleep, Pre-Work, and Post-Work. Participants were instructed to complete the Pre- and Post-Sleep sessions within ~15 min of going to bed and waking up, respectively. In addition, on work days, participants were instructed to complete the Pre- and Post-Work sessions within ~1 hr of "check-in" and "check-out" (i.e., the beginning and end of the entire duty day, respectively). Each test session began with a 5-min touchscreen-based Psychomotor Vigilance Test (PVT), programmed under the same parameters as the Palm-based PVT previously developed at the Walter Reed Army Institute for Research (Lamond, Dawson & Roach 2005; Thorne, Johnson, Redmond, Sing, Belenky & Shapiro 2005) and effectively utilized for various field studies in 24-hr operational environments (Lamond, Petrilli, Dawson & Roach 2006; Ferguson, Lamond, Kandelaars, Jay & Dawson 2008).

Each participant contributed data every day, as described above, for a continuous 3 to 4-week study period. To maintain consistency across days, locations, and conditions, participants were instructed to conduct their test sessions in a comfortable, normally lit environment with as few sensory distractions as possible. Participants were also informed that their safety and professional duties superseded our study requirements, and they were explicitly instructed to not engage in any research activities while actively engaged in or responsible for any work-related activities.

Data Processing and Analysis

Modeling input and predictions. For each individual participant, actigraphy-derived sleep/wake patterns and log data (including activity, time, and location) were merged into a single file suitable for entry into the SAFTE-based Fatigue Avoidance Scheduling Tool (FAST) software package. Actigraphy-based sleep episodes took precedence over manually logged sleep episodes; however, logged sleep was used during periods for which valid actigraphy data were unavailable. To ensure that all participants were modeled with an equally full sleep reservoir at the beginning of the study, a 3-day period of 8-hr sleep at home between 2300-0700 hr was inserted into each individual's schedule prior to the empirical actigraphy and schedule data. Each participant's FAST file was manually checked against his/her processed actigraphy file (to confirm correct sleep/wake patterns) and PDA activity log (to confirm correct times, locations, and sleep periods). Once completed, validated, and entered, the SAFTE-FAST program processed each individual's final composite file to produce a continuous record of model-predicted effectiveness in 30-min increments throughout the study period.

Neurobehavioral performance. Each PVT test yields a number of output variables per session, including mean Reaction Time (RT, msec), mean Speed (1/RT), total Lapses (RTs > 500 msec), and total False Starts (FS, premature responses), all of which were included as objective neurobehavioral performance metrics. However, since the SAFTE model's output metric of Performance Effectiveness is based on PVT speed, we used that metric as the foundation for initial data processing. Specifically, the project database began with a total of 11,567 individual PVT test sessions. Analysis of mean speeds from all sessions revealed a bi-modal distribution, with values ranging from 0.93 to 92.21 per session (mean $\pm$ SD = 4.87 $\pm$ 4.68; median $\pm$ IQR = 3.82 $\pm$ 1.02), high kurtosis (46.68) from the low end of the distribution where most sessions fell, and a very heavy positive skew (5.45). To avoid undue influence of extreme outliers on the eventual calculation of PVT speeds as a percentage of individual baselines, sessions with mean speeds greater than two standard deviations above the grand distribution mean were excluded from further processing. This 6.85% reduction in the database then left 10,775 individual PVT sessions with which to work. We then removed practice sessions recorded during training, sessions with timestamps dated outside the respective individual's activity log, and all sessions from individuals for whom valid FAST reports could not be produced due to corrupted files, processing errors, or unreliable activity logging. Following these corrective procedures, the final dataset was comprised of 10,659 individual PVT sessions from 178 flight attendants (mean $\pm$ SD sessions/participant = 60 $\pm$ 15, range = 15-92).

Once the final PVT database was established, optimum baseline performances were calculated. Defining "baseline" in the types of controlled laboratory studies upon which the SAFTE model was based is relatively straightforward, typically relying on test sessions conducted following several days of optimal sleep conditions but immediately prior to the experimental sleep restriction protocol. Since these orderly sequential conditions do not apply to observational field studies, we developed an analogous performance-based, rather than time-based, method for defining baseline. Specifically, for each individual, we rank-ordered all PVT sessions by mean speed per session, then used the median of the top 10% highest speeds as that individual's baseline. This metric cannot assume that the individual is "well-rested," as in a laboratory study baseline, but like a laboratory study,

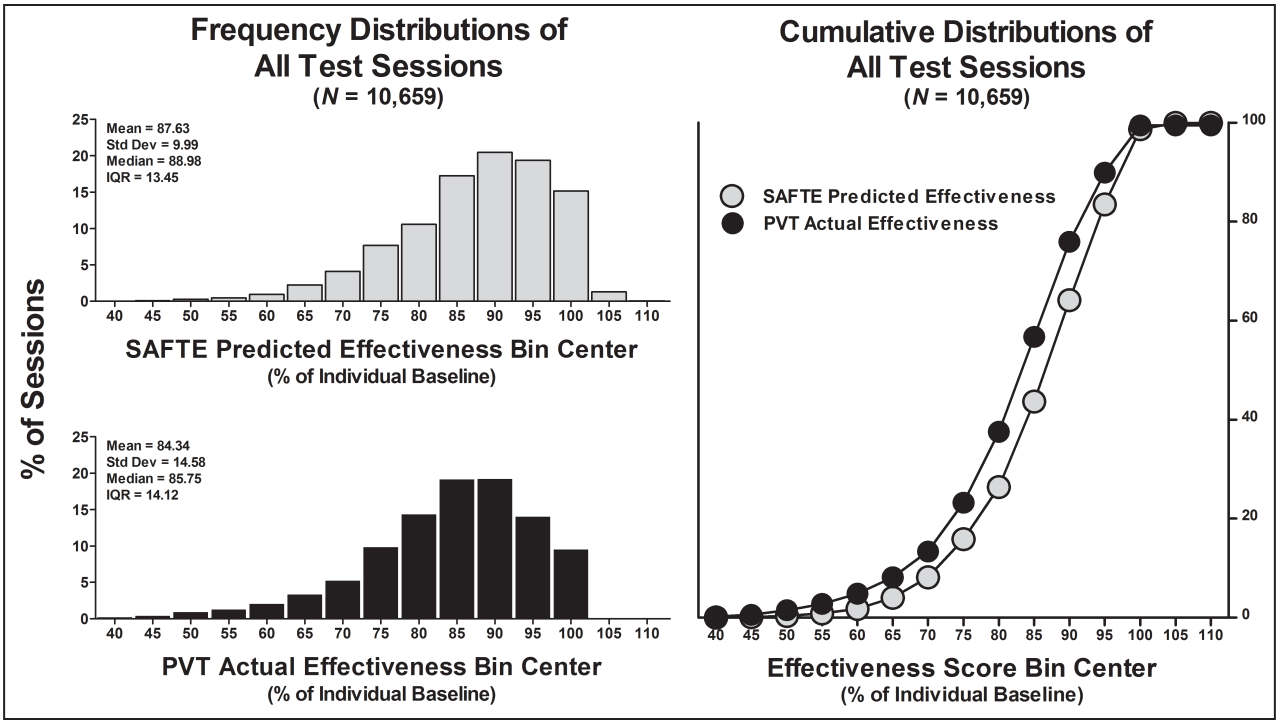

Figure 3: Frequency and Cumulative Distributions of Predicted and Actual Effectiveness Scores for All Test Sessions

this baseline still represents "typical best" performance with ample room for fatigue-induced decrements and countermeasure-induced improvements. Once established, mean speeds for all individual PVT test sessions were then expressed as a percentage of that median of the top 10% fastest speeds. The final outcome metric is comparable to the predicted effectiveness score used by the SAFTE model and is referred to henceforth as "PVT Actual Effectiveness," separate from RT, non-transformed Speed, Lapses, and FS.

Data analysis. Each PVT test session was paired with its corresponding performance effectiveness prediction to the nearest 30-min interval from the respective participant's SAFTE-FAST file. All test session results were then organized into 5% SAFTE Predicted Effectiveness bins (<65%, 65-70%, 70-75%, 75-80%, 80-85%, 85-90%, 90-95%, 95-100%, >100%), and the relationship between mean SAFTE prediction and mean PVT performance across bins was quantitatively assessed via linear regression analysis. Identical regression analyses were performed on all performance metrics, including PVT Actual Effectiveness, RT, Speed, Lapses, and FS. This suite of analyses was also conducted in a nested fashion with increasing operational focus, first with all 10,659 sessions collected throughout the entire study, then only with the 7,533 sessions taken during multi-day work trips (limiting the data to more controlled settings governed by work schedules), then finally separate analyses of only the Pre-Work (n = 1,712) and Post-Work (n = 1,934) sessions to focus on the model's ability to predict variations in performance capacity specifically before and after a work day. Unless otherwise noted, all data are presented as mean ± SEM. All analyses were two-tailed as applicable, and statistical significance was set at $\alpha = .05$.

RESULTS

All Test Sessions

Figure 3 (left panel) shows that the frequency distributions of SAFTE Predicted Effectiveness and the primary outcome measure of PVT Actual Effectiveness were similar in shape, both with a negative skew, as may be expected from measures expressed as a percentage (scaled from zero to ~100). Figure 3 (right panel) reveals a modest gap in cumulative distributions, indicating more PVT Actual Effectiveness data falling within the 75-90% range compared to the model predictions, although both distributions assumed similar, parallel sigmoidal shapes.

As illustrated in Figure 4, linear regression analyses of mean PVT performances across the 5% SAFTE prediction bins revealed significant correlations between SAFTE-Predicted Effectiveness and PVT Actual Effectiveness (R^2 = 0.884, $p < .001$), RT (R^2 = 0.745, $p < .01$), and Lapses (R^2 = 0.486, $p < .05$; all other R^2s < 0.197, ps > .20).

Work Trip Test Sessions

Figure 5 shows that focusing only on sessions completed while participants were away on a work trip had no obvious effect on the frequency distributions or cumulative distributions of SAFTE-Predicted Effectiveness and the primary outcome measure of PVT Actual Effectiveness.

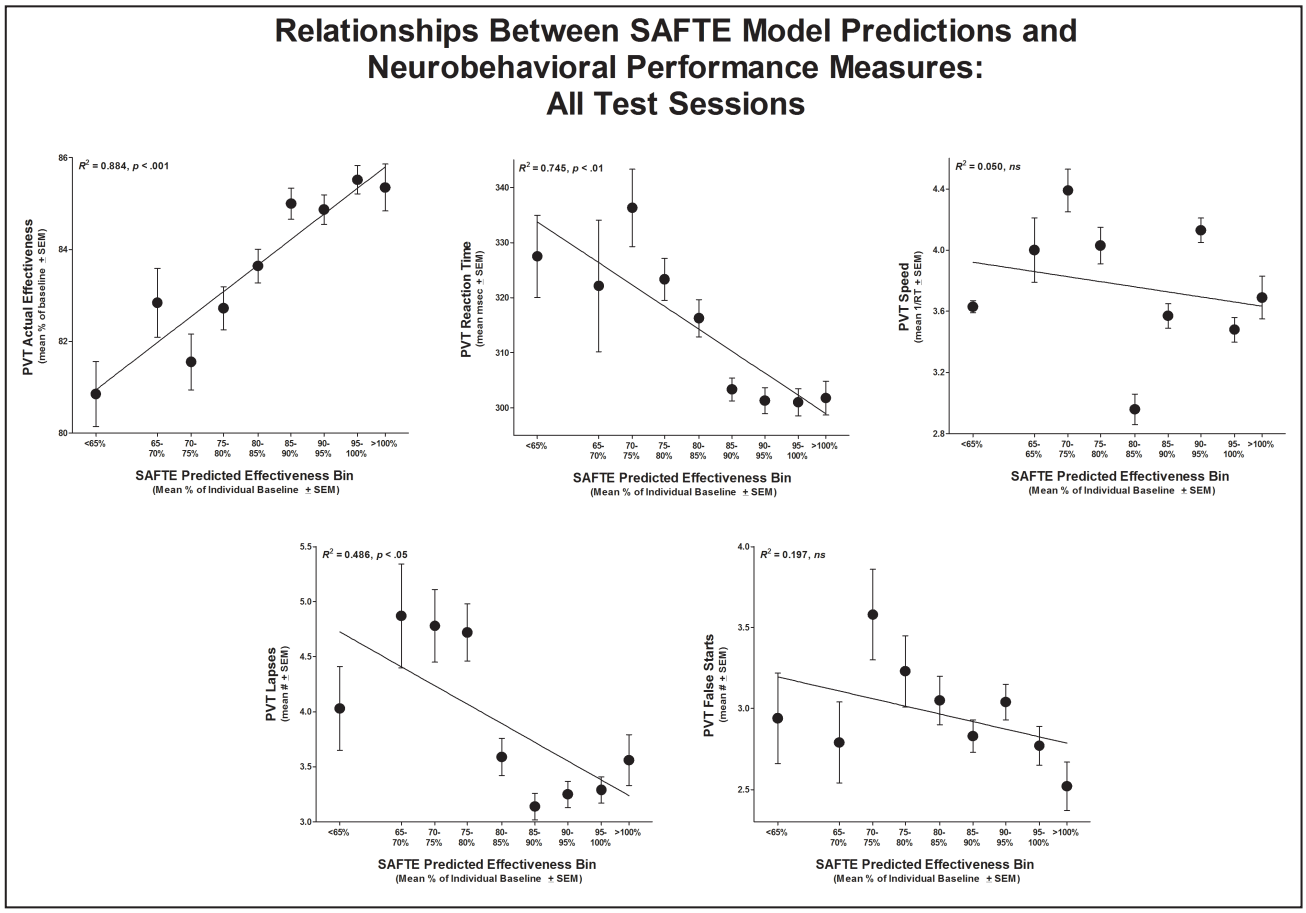

Figure 4: Relationships Between SAFTE Model Predicted Effectiveness and Mean PVT Performances in All Test Sessions

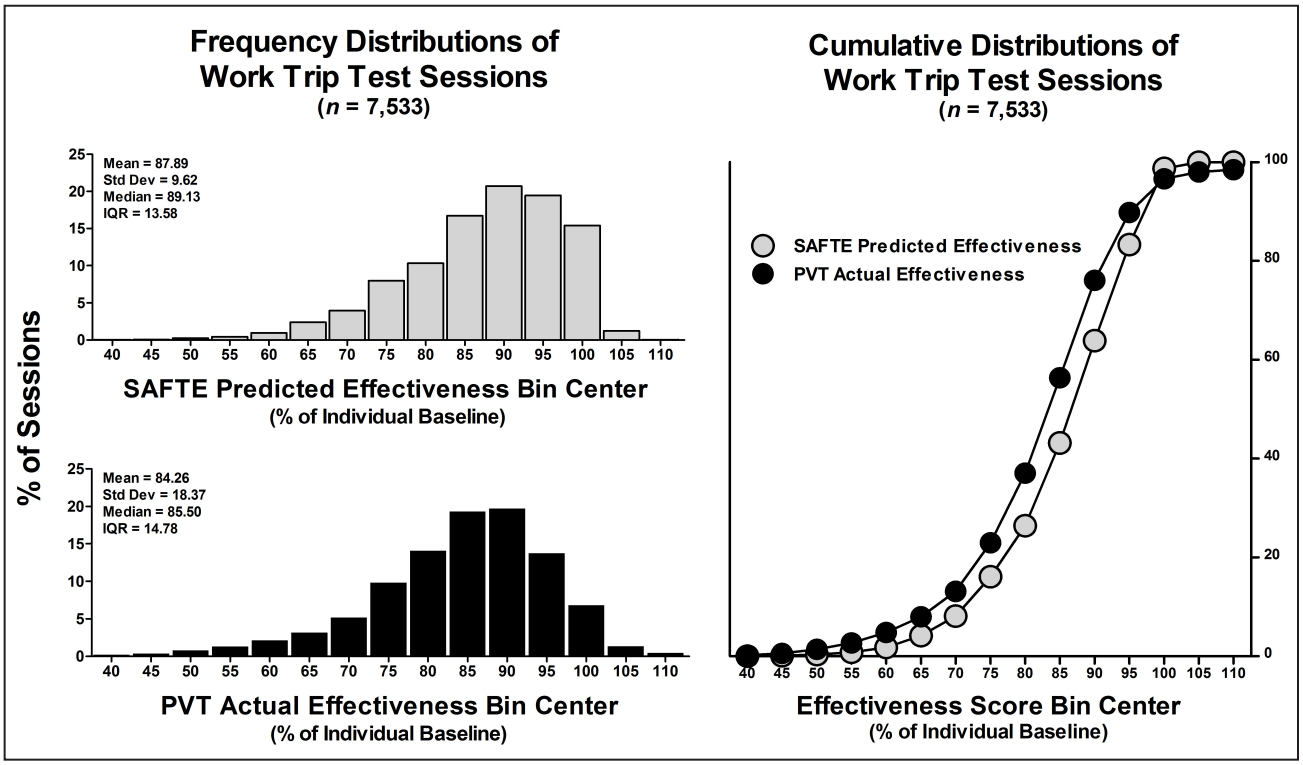

Figure 5: Frequency and Cumulative Distributions of Predicted and Actual Effectiveness Scores for Work Trip Test Sessions

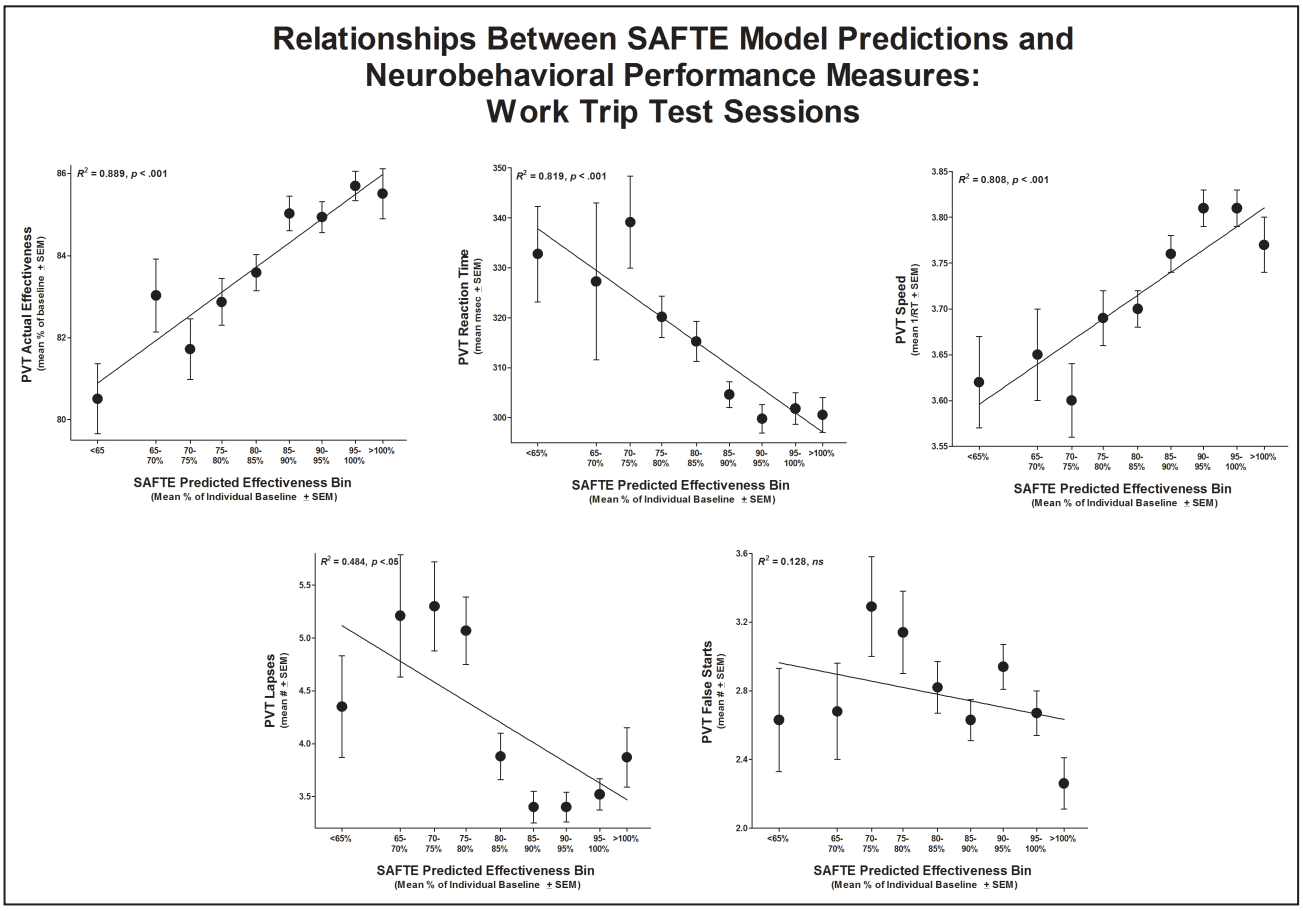

Figure 6: Relationships Between SAFTE Model Predicted Effectiveness and Mean PVT Performances in Work Trip Test Sessions

As illustrated in Figure 6, linear regression analyses of mean PVT performances across the 5% SAFTE prediction bins revealed significant correlations between SAFTE-Predicted Effectiveness and PVT Actual Effectiveness (R^2 = 0.889, p < .001), RT (R^2 = 0.819, p < .001), Speed (R^2 = 0.808, p < .001), and Lapses (R^2 = 0.484, p < .05; FS R^2 = 0.128, p > .30).

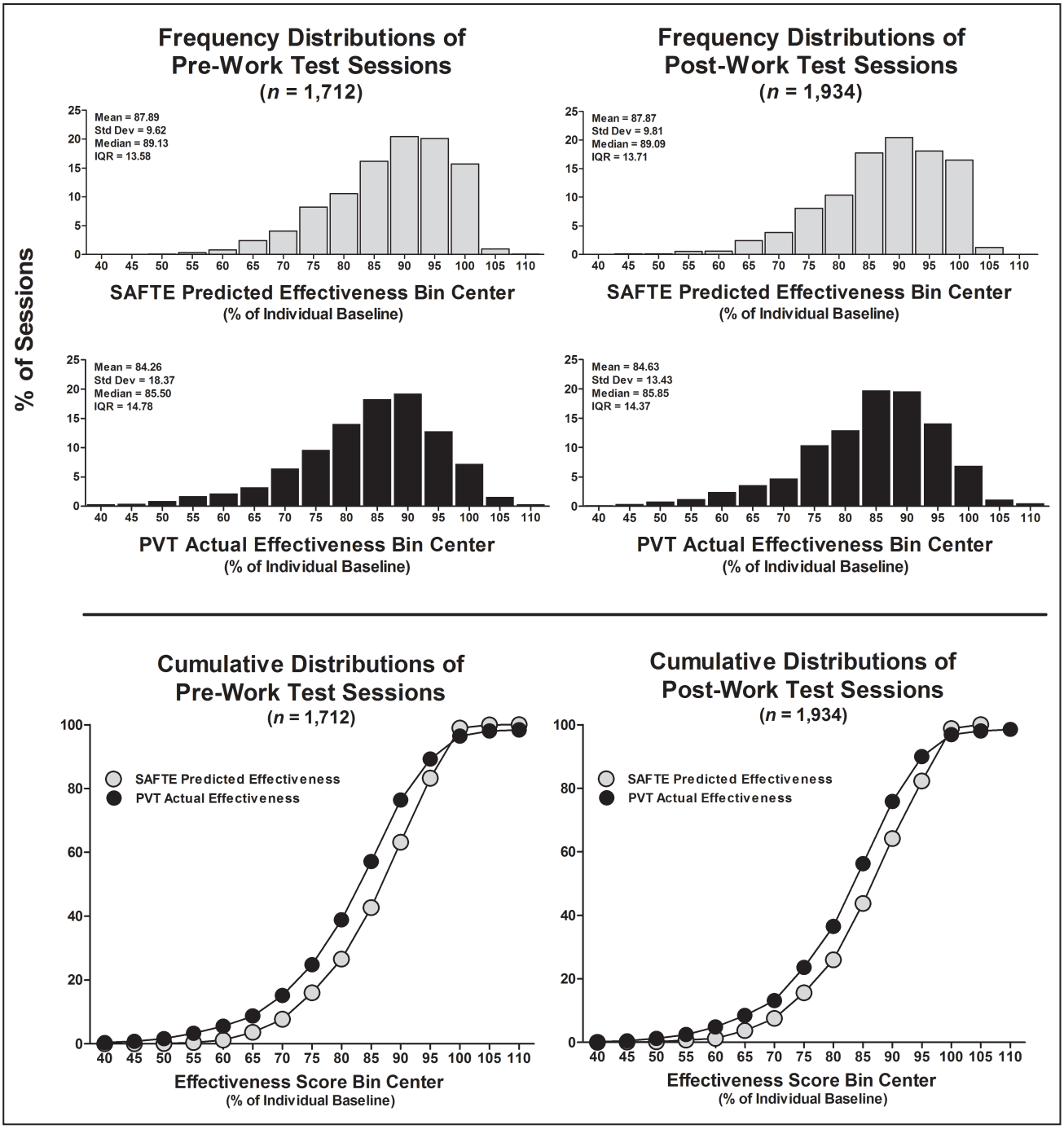

Figure 7: Frequency and Cumulative Distributions of Predicted and Actual Effectiveness Scores for Pre-Work and Post-Work Test Sessions

Pre-Work and Post-Work Test Sessions

Figure 7 shows that focusing on the sessions completed immediately before or after a work day had no obvious effect on the frequency distributions or cumulative distributions of SAFTE-Predicted Effectiveness and the primary outcome measure of PVT Actual Effectiveness.

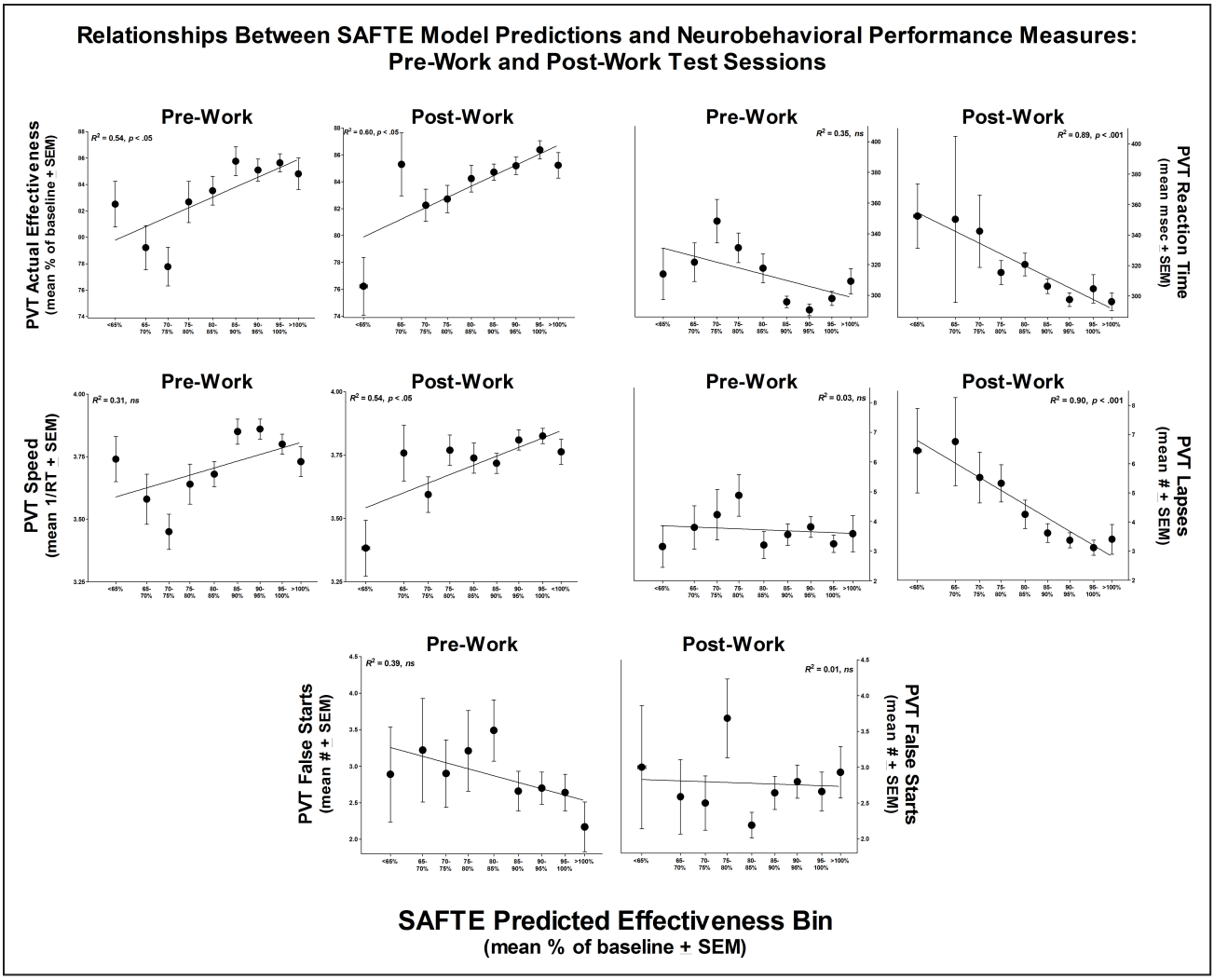

Figure 8: Relationships Between SAFTE Model Predicted Effectiveness and Mean PVT Performances in Pre-Work and Post-Work Test Sessions

As illustrated in Figure 8, linear regression analyses of mean PVT performances across the 5% SAFTE prediction bins reveals a significant Pre-Work correlation between SAFTE-Predicted Effectiveness and PVT Actual Effectiveness (R^2 = 0.530, p < .05; all other Pre-Work R^2s < 0.392, ps > .07). Analysis of Post-Work sessions revealed significant correlations between SAFTE-Predicted Effectiveness and PVT Actual Effectiveness (R^2 = 0.600, p < .05), RT (R^2 = 0.887, p < .001), Speed (R^2 = 0.539, p < .05), and Lapses (R^2 = 0.901, p < .001). Analysis of FS was not significant (R^2 = 0.006, p > .80).

DISCUSSION

Whether examining all test sessions, sessions completed throughout a multi-day work trip, or just those sessions completed before and after a work day, predicted performance effectiveness scores rendered by the SAFTE model correlated significantly with average performances on multiple PVT metrics. Specifically, as predicted effectiveness decreased, RTs increased, Speed decreased, and Lapses increased—all patterns consistent with impaired neurobehavioral performance capacity. Importantly, SAFTE-Predicted Effectiveness most strongly and consistently correlated (positively) with the analogous PVT Actual Effectiveness metric, and the SAFTE model's predictive ability generally increased with increasing focus on test sessions whose timing was governed by operational schedules. Given the broadly representative sample of participants, extensive longitudinal and standardized data collection, and the use of actual sleep/wake/work patterns, the results strongly support the validity of the SAFTE model for predicting population-level variations in objective performance effectiveness.

Of course, for as encouraging as these results are, there are several features of the dataset worth considering when interpreting the findings. First, despite the strong correlations between SAFTE-Predicted Effectiveness and PVT Actual Effectiveness, the concordance between the two variables was limited by the differences in the range of the two metrics. Mean predicted effectiveness values ranged from well below 65% to above 100%, whereas the paired actual mean effectiveness scores ranged from 75-90% (evident in Figures 4, 6, and 8).

It is in this context that the differences between the laboratory data used to develop the SAFTE model and the field methods used in the present study may be most relevant. For example, the use of a PDA-based touchscreen PVT vs. the traditional push-button PVT "box," the use of a 5-min PVT session vs. the traditional 10-min session, and the technical limitations of off-the-shelf consumer-grade electronics may all have contributed to limit the sensitivity, or at least the functional range, of mean performance effectiveness. In addition, the differences between laboratory- and field-based definitions of "baseline" may also have affected the final calculations, although the influence of this factor is virtually impossible to determine.

Nonetheless, these concerns only apply to the PVT Actual Effectiveness variable since it is analogous to the SAFTE-Predicted Effectiveness metric. We contend that differences in scale between predicted and actual effectiveness are less important than the essential finding that periods of relatively high- and low-predicted performance capacity were strongly associated with respective periods of relatively high and low actual performance capacity, as measured by several PVT variables. Since the 5-min PVT itself is not a task inherent to aviation operations but rather measures core neurobehavioral processes necessary for more complex operational tasks (Lim & Dinges, 2008), the key to the SAFTE model's validation is that when the model predicts peak performance, one is most likely to be at his or her best, and when the model predicts severely impaired performance, one is most likely to be at their worst, regardless of how "best" and "worst" are quantified or otherwise translated to the operational context.

Another noteworthy feature of the data was the high variability in mean PVT performances observed at the low range of predicted effectiveness, particularly at and below the 75% bin. The reasons for this are unclear, although we do not necessarily view this pattern observed in nearly all outcome variables as a limitation. One possibility is that fewer test sessions were available in the low-range bins, and hence higher SEMs after averaging; however, re-analysis of the data in bins of 100 sessions each still yielded a similar pattern (results not shown).

More likely explanations draw from specific features of the fatigue construct itself. One possibility comes from an emerging area in fatigue research on inherent individual differences in sleep need and vulnerability to fatigue (Goel & Dinges, 2011; Van Dongen, Baynard, Maislin & Dinges 2004a). Simply put, not every individual will exhibit performance impairments or the same level of impairment under sleep/wake/work patterns expected to produce fatigue in the general population, and conversely, particularly vulnerable individuals may consistently implement prophylactics and countermeasures (e.g., caffeine, nicotine, light exposure, exercise) to mitigate fatigue effects, regardless of their schedules. In both cases, individuals who perform with high intra-individual consistency, despite model predictions, will add variability to the average performances provided by their more susceptible colleagues.

Another potential contributor is the disparity between subjective fatigue and objective decrements in performance capacity (Van Dongen, Maislin & Dinges 2004b). Indeed, one particularly insidious feature of fatigue is a reduced ability to recognize the transition from baseline to moderate impairment, so individuals whose performance capacity is altered by fatigue may not realize it at predicted effectiveness levels above 75%, thereby yielding objective performance outcomes in line with model predictions. Yet, sleep/wake/work patterns that modeled as extremely impaired may have produced sufficient subjective fatigue to provoke countermeasure implementation (which we did not monitor or control), thus mitigating the objective performance decrements

one would observe in a controlled laboratory setting, ultimately producing PVT performances more similar to sessions associated with higher predicted effectiveness bins.

Finally, another likely contributor is the notion of fatigue as "state instability" (or state lability; Dinges & Kribbs, 1991; Dorrian, Rogers & Dinges 2005), which defines fatigue more as inconsistent performance while the brain struggles to maintain vigilance, rather than consistently suppressed performance, reflecting the steady state of sleep pressure. From this perspective, higher variability at the very low end of predicted effectiveness would be expected, especially when coupled with individual differences in sensitivity to fatigue or other extraneous factors, and our ability to detect this variability actually speaks well of the 5-min touchscreen PVT's sensitivity as a field research tool (cf. Lamond et al., 2006; Ferguson et al., 2008). Since this study was intentionally designed to capture naturally occurring sleep/wake/work patterns and behavior without any field-validated means of quantifying individual vulnerability to fatigue, we must accept all of the possibilities described above as potential complications.

Nonetheless, it is at least provocative, if not encouraging from a model validation perspective, to observe that this apparent 75% predicted effectiveness "cutoff" point in performance stability is nearly identical to the point at which accident severity risk increases 5-fold in freight rail operations (77%; Hursh et al., 2011). Despite the various issues described above, the emergence of significant orderly relationships between model predictions and multiple objective neurobehavioral performance metrics further supports the SAFTE model's general validity for use in 24-hr operational settings while providing direct support for the model's applicability to commercial aviation.

Regarding future directions, our primary Flight Attendant Field Study report (Roma et al., 2010) was based on the same PVT performance data utilized for the present study's modeling analysis and revealed pervasive fatigue manifested as significant performance decrements in all cabin crew at the start of their work shifts relative to baseline. Performance capacity worsened from Pre-Work to Post-Work as expected, but with differential effects based on the broad demographic factors of Carrier Type, Seniority, and Flight Ops. If one accepts the validity of the SAFTE model for predicting risk as demonstrated in rail operations and supported by the present study, then a worthy flight attendant-specific follow-up analysis would be a detailed model-based investigation of risk as a function of these demographic variables (e.g., percentage of duty time spent below various predicted effectiveness criteria). Since the SAFTE model incorporates circadian and homeostatic components of clear relevance to aviation, regardless of demographics, such an analysis could yield valuable first insights on the operational variables underlying the performance differences observed between the various flight attendant groups, which could warrant more detailed analyses to empirically inform decision-making by regulatory agencies, labor unions, airline management, and other organizations with a vested interest in cabin safety.

Broader implications of the present work relate to further model development and application. For example, a unique feature of the SAFTE-FAST system is the "Auto-Sleep" function, which estimates sleep duration in the absence of empirical input. The Auto-Sleep function was not used in this study because actual sleep measurements were available; yet, the extensive sleep data available from this study could be used to inform the Auto-Sleep function in SAFTE-FAST. Consistent with the evidence-based development approach of the SAFTE model it serves, Auto-Sleep's parameters were built on the empirical sleep/wake patterns of shiftworking rail employees (Federal Railroad Administration Research Results, 2011; Pollard, 1991). The extensive compendium of objective and subjective sleep/wake/work pattern data from the present study could therefore be used to validate and/or calibrate the current Auto-Sleep function specifically for commercial aviation.

Another emerging issue in fatigue modeling is individualization, i.e., incorporating flexible parameters based on trait-like individual differences in response to the various fatigue-producing inputs accounted for by any given model (CASA, 2010; Van Dongen, Bender & Dinges 2012). Most model predictions represent a population average, and the SAFTE model is no exception, although transforming validation data to a percentage of individual baselines accommodates individual differences to some extent. However, by its very nature, such a transformation only accounts for differences in some baseline parameter derived from the data post-hoc and does not accommodate inherent differences in the extent to which individuals are vulnerable to fatigue, for example, via differences in sleep need, sleep inertia, or circadian phasing and amplitude.

As we have seen in the present study and others (Roma et al., 2010; Greeley et al., in press), appropriately defining individual baselines in such a way that is conceptually valid, statistically beneficial, and operationally feasible is a complex matter with no simple solutions, even with a large post-hoc dataset with which to work. These issues would only be exacerbated by the need to develop *a priori* models at the individual level, especially if intended for field application in very large and exceptionally mobile populations such as commercial aviation. But this is a

challenge the scientific and operational communities must eventually confront to maximize the benefits of biomathematical modeling as a fatigue risk management tool.

In conclusion, predictive fatigue modeling for operational use is still a relatively young science, so all theoretical and empirical work in this area make important contributions, nonetheless. The present study utilized actual sleep/wake/work data from a broadly representative sample of professional cabin crew to demonstrate clear relationships between performance effectiveness predicted by the SAFTE model and objective performance outcomes in the field. Despite the study's limitations, the data presented herein further support the predictive validity of the SAFTE model, and specifically support the model's validity within the exceptionally dynamic operational environment of commercial aviation.

In terms of vulnerability to fatigue, we believe it is reasonable to assume that the professional cabin crew population is not inherently different at the genetic/biological level than any other sub-group within the aviation community. Similarly, it is reasonable to assume that the commercial aviation population is not inherently different than any other group of generally healthy adults exposed to round-the-clock work schedules. If so, then the SAFTE model and the present study's comprehensive dataset are valuable resources that could continue to generate important insights on sleep/work/wake patterns and neurobehavioral performance capacity in the "real world." As such, we encourage continued investigation of the Flight Attendant Field Study database and further development of the SAFTE model in the spirit of science-based technologies for improving the safety, performance, health, and quality of life of those who work in and rely on 24-hr operations.

REFERENCES

Åkerstedt, T. (1995). Work hours and sleepiness. *Neurophysiologie Clinique*, *25*, 367-375.

Avers, K.B., King, S.J., Nesthus, T.E., Thomas, S., & Banks, J. (2009). *Flight attendant fatigue, Part I: National duty, rest, and fatigue survey*. (Report No. DOT/FAA/AM-09/24). Washington, DC: Office of Aerospace Medicine, Federal Aviation Administration.

Basner, M., & Dinges, D.F. (2011). Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, *34*(5), 581-591.

Civil Aviation Safety Authority (CASA), Human Factors Section. (2010). *Biomathematical fatigue modelling in civil aviation fatigue risk management*. Canberra, Australia: Civil Aviation Safety Authority. [http://www.casa.gov.au/wcmswr/_assets/main/aoc/fatigue/fatigue_modelling.pdf]

Dinges, D.F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(suppl.2), 4-14.

Dinges, D.F., & Kribbs, N.B. (1991). Performing while sleepy: Effects of experimentally induced sleepiness. In: T.H. Monk (Ed.), *Sleep, Sleepiness and Performance* (pp. 97–128). Chichester, England: Wiley.

Dorrian, J., Rogers, N.L., & Dinges, D.F. (2005). Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In Kushida, C.A. (Ed.), *Sleep Deprivation: Clinical Issues, Pharmacology and Sleep Loss Effects* (pp. 39-70). New York, NY: Marcel Dekker, Inc.

Federal Railroad Administration Research Results. (2011). *Measurement and estimation of sleep in railroad worker employees* (Report No. DOT/FRA/RR11-02). Washington, DC: U.S. Federal Railroad Administration, Department of Transportation.

Ferguson, S.A., Lamond, N., Kandelaars, K., Jay, S.M., & Dawson, D. (2008). The impact of short, irregular sleep opportunities at sea on the alertness of marine pilots working extended hours. *Chronobiology International*, *25*(2), 399-411.

Goel, N., & Dinges, D.F. (2011). Behavioral and genetic markers of sleepiness. *Journal of Clinical Sleep Medicine, 7*(5 Suppl), S19-21.

Greeley, H.P., Roma, P.G., Mallis, M.M., Hursh, S.R., Mead, A.M., & Nesthus, T.E. (in press). *Field study evaluation of cepstrum coefficient speech analysis for fatigue in aviation cabin crew* (Report No. pending). Washington, DC: Office of Aerospace Medicine, Federal Aviation Administration.

Hursh, S.R., Fanzone, J.F., & Raslear, T.G. (2011). *Analysis of the relationship between operator effectiveness measures and economic impacts of rail accidents* (Report No. DOT/FRA/ORD-11/13). Washington, DC: U.S. Federal Railroad Administration, Department of Transportation.

Hursh, S.R., Redmond, D.P., Johnson, M.L., Thorne, D.R., Belenky, G., Balkin, T.J., Storm, W.F., Miller, J.C., & Eddy, D.R. (2004). Fatigue models for applied research in warfighting. *Aviation, Space, and Environmental Medicine*, *75*(3,Suppl.):A44–53.

Hursh, S.R., Raslear, T.G., Kaye, A.S., & Fanzone, J.F. (2008). *Validation and calibration of a fatigue assessment tool for railroad work schedules* (Report No. DOT/FRA/ORD-08/04). Washington, DC: U.S. Federal Railroad Administration, Department of Transportation.

Hursh S.R., & Van Dongen, H.P.A. (2010). Fatigue and performance modeling. In M.H. Kryger, T. Roth & W.C. Dement (Eds.), *Principles and Practice of Sleep Medicine, 5th Ed.* (pp. 745-752). Philadelphia: Elsevier Saunders.

Lamond, N., Dawson, D., & Roach, G.D. (2005). Fatigue assessment in the field: Validation of a hand-held electronic psychomotor vigilance task. *Aviation, Space, and Environmental Medicine*, *76*(5), 486-489.

Lamond, N., Petrilli, R.M., Dawson, D., & Roach, G.D. (2006). Do short international layovers allow sufficient opportunity for pilots to recover? *Chronobiology International*, *23*(6), 1285-94.

Lim, J., & Dinges, D.F. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, 1129, 305-322.

Mallis, M.M., Banks, S., & Dinges, D.F. (2010). Aircrew fatigue, sleep need and circadian rhythmicity. In E. Salas and D. Maurino (Eds.), *Human Factors in Aviation, 2nd Edition* (pp. 401-436). Burlington, MA: Academic Press.

Nesthus, T.E., Schroeder, D.J., Connors, M.M., Rentmeister-Bryant, H.K., & DeRoshia, C.W. (2007). *Flight attendant fatigue*. (Report No. DOT/FAA/AAM-07/21). Washington, DC: Office of Aerospace Medicine, Federal Aviation Administration.

Pollard, J. (1991). *Issues in locomotive crew management and scheduling* (Report No. DOT/FRA/RRP-91-01). Washington, DC: U.S. Federal Railroad Administration, U.S. Department of Transportation.

Roma, P.G., Mallis, M.M., Hursh, S.R., Mead, A.M., & Nesthus, T.E. (2010). *Flight attendant fatigue recommendation II: Flight attendant work/rest patterns, alertness, and performance assessment* (Report No. DOT/FAA/AM-10/22). Washington, DC: Office of Aerospace Medicine, Federal Aviation Administration.

Spencer, M.B., & Robertson, K.A. (2007). The application of an alertness model to ultra-long-range civil air operations. *Somnologie*, *11*, 159-166.

Thorne, H., Hampton, S., Morgan, L., Skene, D.J., Arendt, J. (2008). Differences in sleep, light, and circadian phase in offshore 18:00-06:00 h and 19:00-07:00 h shift workers. *Chronobiology International*, *25*(2&3), 225-235.

Thorne, D.R., Johnson, D.E., Redmond, D.P., Sing, H.C., Belenky, G., & Shapiro, J.M. (2005). The Walter Reed palm-held psychomotor vigilance test. *Behavior Research Methods*, *37*(1), 111-118.

Van Dongen, H.P.A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviation, Space, and Environmental Medicine*, *75*(3, Suppl.), A15–36.

Van Dongen, H.P., Baynard, M.D., Maislin, G., & Dinges, D.F. (2004a). Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. *Sleep, 27*(3), 423-433.

Van Dongen, H.P., Bender, A.M., & Dinges, D.F. (2012). Systematic individual differences in sleep homeostatic and circadian rhythm contributions to neurobehavioral impairment during sleep deprivation. *Accident Analysis and Prevention, 45*(Suppl), 11-16.

Van Dongen, H.P.A., Maislin, G., & Dinges, D.F. (2004b). Dealing with interindividual differences in the temporal dynamics of fatigue and performance: Importance and techniques. *Aviation, Space and Environmental Medicine, 75*(3, Suppl.), A147–154.